International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 4, May-June 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Predicting the Level of Crowdfunding Outcome in Africa: A Supervised Machine Learning Approach

Isaac Okyere Paintsil[1], Zhao Xicang[1], Oliver Joseph Abban[2]

[1]School of Finance and Economics, Jiangsu University, Zhenjiang, China
[2]Institute of Applied Systems and Analysis (IASA), School of Mathematical Science, Jiangsu University, Zhenjiang, China

ABSTRACT

One crucial challenge of crowdfunding is that it is hard for fundraisers and backers to anticipate the outcome of crowdfunding campaigns. Across platforms, many crowdfunding campaigns fail to achieve their funding goal. Hence, studies focusing of the outcome of crowdfunding is also growing. In this study, we implement a supervised machine learning methodology to investigate the determinants of the level of crowdfunding with emphasis on Africa. The statistical methods used in the study produced a high prediction accuracy. Irrespective of the method used, the number of backers is identified to be the most important predictor of the level of funding. Also, the average amount pledged to the project and the duration of the project are important features that predict the level of funding.

KEYWORDS: *Crowdfunding, decision tree, supervised machine learning, Africa*

How to cite this paper: Isaac Okyere Paintsil | Zhao Xicang | Oliver Joseph Abban "Predicting the Level of Crowdfunding Outcome in Africa: A Supervised Machine Learning Approach" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4, June 2021, pp.1242-1254, URL: www.ijtsrd.com/papers/ijtsrd42539.pdf

IJTSRD42539

1. INTRODUCTION

The World Bank Group (2015) projects that crowdfunding has a great potential to increase financial access in developing economies including many African countries. However, the crowdfunding phenomenon is still considered at its infancy in Africa and only accounts for a small percentage of the global crowdfunding market volume. According to KPMG (2015), African financial sectors are predominantly underdeveloped and face the challenge of low penetration of traditional financial institutions including banks. By virtue of the vast funding gap existing in the Africa, crowdfunding is believed to possess a huge potential for economic transformation in Africa. This study attempts to investigate the underlying factors that promotes crowdfunding in some countries in Africa.

One crucial challenge ofcrowdfunding is that it is hard for fundraisers and backers to anticipate the outcome of crowdfunding campaigns. Most crowdfunding platforms operate on two models namely, “Keep-it-all” (KIA) (where the fundraiser sets a campaign goal and keeps the entire amount raised irrespective of whether the campaign goal was achieved or not) and “All-or-Nothing” (where the fundraiser keeps nothing unless the campaign goal is achieved). In most situations, many crowdfunding campaigns fail to reach the funding target (Cumming et al., 2014; Mollick, 2014). This phenomenon has attracted many researchers to investigate the determinants of crowdfunding success. For instance, Davies and Giovannetti (2018) suggest that signals relating to a project’s creator experience, previously backed campaigns, early funding, early backing and external social capital have positive effects on the outcome of the crowdfunding campaign. Also, the goal and duration have negative consequences on a campaign outcome. As we can see from Figure 1, about 37.44% of campaign projects on Kickstarter achieved success and approximately 63% failed to reach their campaign goals in the year 2019.

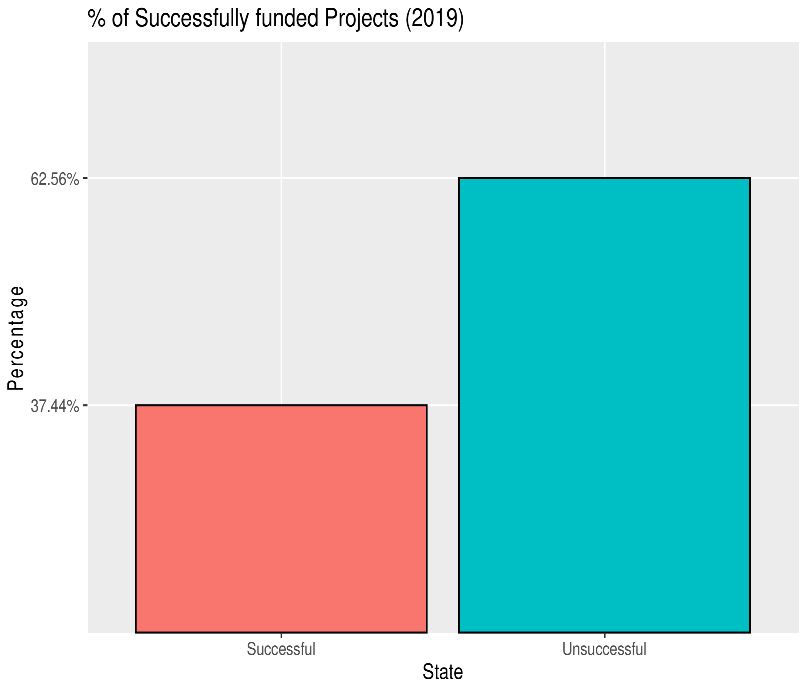

Figure 1Kickstarter % of successfully funded projects in 2019 Source: Statista

2. Literature review

Zheng et al., (2014) argue that successful crowdfunding also depends on what is valued by capital providers. Cholakova and Clarysse (2015) report that the motivation for capital providers to contribute towards crowdfunding is divers including rewards and financial returns, help other people in need and support a cause, and to form relationships and be part of a community (Galuszka & Brzozowska, 2017; Gerber & Hui, 2014; Lam & Law, 2016). Ryu and Kim)2015) classify six incentives that motivates capital providers namely: interest, playfulness, philanthropy, reward, relationship, and recognition. They also identify four categories of crowdfunding sponsors, namely, the angelic backer who characterized by philanthropic motives, the reward hunter who is characterized by reward motives, the avid fan who characterized by several motives except rewards, and the tasteful hermit who are similar to the avid fans but are less driven by recognition and relationship motives). Choy and Schlagwein (2016) also categorized crowdfunding capital providers into four categories namely: the intrinsic-individual, the intrinsic-social, the extrinsic-individual, and the extrinsic-social motivations. On another hand, Gerber and Hui (2014) explain that distrust for a fundraiser's management of funds often discourages capital providers. This situation is often associated with platforms where the fundraiser full access to the funds even when the campaign target has not been met.

Ciuchta et al., (2016) identify two categories of capital providers namely, prevention-focused funders and promotion focused funders. They argue that prevention-focused funders often look out for negative feedbacks whereas promotion focused funders look out for positive feedbacks to inform their contribution decisions. Furthermore, promotion focused funders are more responsive to social information in crowdfunding campaigns. Burtch et al., (2013) explain that capital providers are often inclined towards culturally similar and geographically proximate fundraisers referred to as "home bias" by Lin and Viswanathan (2016). By the same token, Mollick (2014) suggest that geography plays an important role in the success of crowdfunding campaigns, since a proportionately greater creative population in a fundraiser's city is associated with a greater probability of success. Agrawal et al., (2015b) identify disparities in funding patterns between local and distant capital providers. They argue that local capital providers appear less responsive to information about the cumulative funds raised by the fundraiser. Galuszka and Brzozowska (2016) show that capital providers are willing to support projects from friends and local fundraisers.

Ordanini et al., (2011) stress that crowdfunding platforms play different roles depending on the type of crowdfunding they support. Platform design enables the capital providers to grasp campaign messages (Choy & Schlagwein, 2016). The policies and standards which guide the platform contribute to determining the probability of success. For instance, Farnel (2014), reveal that the Indiegogo's guidelines and standards were more welcoming to crowdfunding gender or sexual reassignment surgery projects than Kickstarter and YouCaring platforms. Yuan et al., (2016) explain that platform cultures may are best comprehended by performing a semantic analysis of campaigns on a platform to uncover the topical features (i.e., latent semantics) of successful campaigns. The availability of other projects on the platform has an impact on the probability of success. Meer (2014) and Parker (2014) explain that a handful of good projects competing with a particular can cause the project to be funded. Parker (2014) argue that not every capital provider makes informed investment decision and thus most capital providers tend to follow few informed investors. Hence, capital providers tend to look out for other people who often go back to contribute toward good projects.

3. Supervised machine learning

Machine learning is a subfield of artificial intelligence that learns patterns in data to perform specific tasks(Lopez de Prado, 2018). The main classes of machine learning are supervised machine learning, unsupervised machine learning and reinforcement learning. In supervised machine learning, computers learn from a labeled data to predict a target variable. Supervised machine learning algorithms

basically perform two different types of tasks namely, regression and classification.

Supervised machine learning approach is gaining popularity for predicting crowdfunding outcome. Ren et al., (2018) uses supervised machine learning methods to predict the daily funding of crowdfunding projects. Similarly, Yu et al., (2018) adopts supervised machine learning approach to predict the success of a crowdfunding campaign. Ahmad et al., (2017)uses random forest model with optimally weighted classifiers to predict crowdfunding success. Also, Chung and Lee (2015)examine the impact of a project's temporal features to predict the range of pledges.

Algorithms explored in the study

3.1. Decision tree

Decision tree, introduced in the 1960's is commonly used in data mining for establishing classification systems based on multiple covariates and for developing prediction algorithms for a target variable (Ren et al., 2018). A Decision tree analysis takes a graphical representation of diverse solutions format. The process forms a flowchart structure made of nodes, branches and leaves building blocks. These features are explained as follows:

1. The root node which is the first node that decides the entire sample space.
2. The splitting which are the branches that emanates from the root node and also represents a set of decision alternatives where one and only one decision can be selected.
3. Internal nodes which represent the probability that an outcome from the branch will occur.
4. Leaf representing the terminal node that predicts the final outcome following a path from the root node.

Decision tree uses attributes that has the highest information gain. The information gain is determined by decrease in entropy after the data is split on an attribute. Entropy is basically the measure of information. Below is a mathematical representation of Entropy. The entropy equation is as below:

$$E(p) = -\sum_{i=1}^{C} p_i \log_2 p_i \qquad 1$$

where p is the data set, C is a dataset with classes and p_i represent the frequency of class i in the data set. The total entropy for each split is computed and subtracted from the entropy before the split to derive the information gain represented in the equation below:

$$information\ Gain = Entropy\ (before) - \sum_{j=1}^{K} Entropy(j, after) \qquad 2$$

The attribute with the highest information gain becomes the decision node for a repeated process to arrive at an entropy of 0 which is also known as the leaf node.

We also explored some ensemble learning algorithms. Ensemble learning in machine learning refers to an assembly of models that learn a target function by training a number of individual learners. Ensemble learning provides high accuracy and efficiency using various methodologies including bagging, boosting, and stacking.

3.2. Random forest

Breiman's Random forest (Breiman, 2001) algorithm is an ensemble of regression trees which develop easy to visualize decision rules for classification and regression. More formally random forest can be expressed as:

$$g(x) = f_0(x) + f_1(x) + f_2(x)\ldots \qquad 3$$

where the function $g(x)$ is the sum of simple base models and $f_i(x)$ is a simple decision tree. The Random forest algorithm builds an ensemble of trees and each tree gives a classification or class votes. Each tree within the random forest model is trained from a random subset of data from a training data set thus the randomness of the model is implied. Random forest combines the output of multiple decisions trees to arrive at a final output. The final output is either determined by the averaging of the number of decision trees or by the majority voting from each tree.

3.3. Extreme gradient boosting

Gradient boosting algorithm introduced by Tianqi Chen in 2014 is mostly used in machine learning are a result of the need to improve upon weak classifiers and improve prediction accuracy(see Chen et al., 2020). Boosting technique is an ensemble approach the combines various weak learners to build a high precision strong learner. Boosting is also very useful for large data sets and also for analysis that require high prediction accuracy. Furthermore, boosting is extremely useful for classification, feature selection, and multiclass categorization. extreme gradient boost is characterized by high performance, speed and scalable in nature. The algorithm can also handle parallel processing and control for overfitting through its regularization, sparsity awareness, weighted quantile sketch and inbuilt cross validation.

4. Materials

The data for this study is taken from The Crowdfunding Data Center. The Crowdfunding Data Center specializes in tracking and providing information on various crowdfunding campaigns from around the world. Data on crowdfunding projects involving African countries that ended between 1st September 2016 and 30th September 2016 are collected for the analysis. Table1 displays countries, the number of campaigns, and crowdfunding platforms for campaign projects that ended within the period. Overall, 237 crowdfunding projects on Africa launched on Kickstarter.com, Indeigogo.com, and Fundraiser.com are reported within the study period. Kickstarter and Indiegogo are among the leading non-equity-based crowdfunding platforms globally (Steinberg 2012, Mollick 2015). Out of the total number of crowdfunding projects in Africa sample, 11 projects were launched on Kickstarter.com platform whereas 221 projects were launched on indeigogo.com and 5 projects were launched on fundrazrs.com. The steps involved in launching a product includes a registration to join the community which also involves providing personal details to the platform providers. Then the fundraiser receives the clearance to design and promote a webpage with details of the crowdfunding project to attract the crowd to make a decision to contribute towards the project or not.

Table 2 presents the structure of the data collected for the study. The data includes the identification number, campaign title, category, type of crowdfunding country, a link to the campaign's URL, launch date, end date, flexible fund, fundraiser's name, fundraiser's email, social network sites, project website, campaign target, the amount pledged, and number of backers. The feature we incorporate in the study include the availability of website (WBS), average amount

pledged to the project (AVP), and the availability of the fundraiser's personal information (INFO). For cognitive social capital we include the type of crowdfunding (TYP) and the funding duration (DRTN). With respect to structural social capital, we incorporate the number of backers (BKS)and social media account (SMC). The proposed model takes the following structure.

$$FND = f(DRTN, TYP, INFO, BKS, AVP, WBS, SCM)$$

where FND is the crowdfunding campaign level of funding. The analysis is conducted using *R* statistical software version 3.6.1 in an *R* studio version 1.2.1335 (*R* is a free statistical software for computing and graphics).

Table 1List of countries sampled

Country	No. of Projects	Kickstarter	Indeigogo	Fundrazr
Benin	3		3	
Botswana	2	1	1	
Congo DRC	2	1	1	
Cameroon	9		9	
Cape Verde	2	2		
Algeria	1		1	
Egypt	15		15	
Ethiopia	3	1	2	
Ghana	10		10	
Kenya	19		19	
Liberia	1		1	
Morocco	8		7	1
Madagascar	2		1	1
Mali	1	1	1	1
Mauritius	2		2	
Malawi	1		1	
Mozambique	4		4	
Namibia	1		1	
Niger	1	1		
Nigeria	37		37	
Rwanda	3		3	
Senegal	1		1	
Somalia	1		1	
Tunisia	7		7	
Tanzania	9		9	
Uganda	18	1	17	
South Africa	64	2	60	2
Zambia	6	1	5	
Zimbabwe	4		4	
Total	**237**	**11**	**221**	**5**

Table 2Selected attributes from The Crowdfunding Data Center.

Variable	Description	Data Type
ID	Project index number	Letters and numbers
Percentage funded	Percentage of funding target raised over the campaign	Numbers
Date added	The date a specific crowdfunding campaign was launched.	Date
End Date	The date a crowdfunding project was closed on the platform.	
Category	Animals, Arts, Comics, Community, Dance, Design, Education, Environment, Fashion, Film, Food, Gaming, Health, Music, Photography, Politics, Religion, Small business, Sports, Technology, Theater, Transmedia, Video, Writing.	Text
Types	The type of crowdfunding	Text
Platform	The name of the crowdfunding platform	
Link	Link to information on the crowdfunding campaign	Text
Title	The title for a specific crowdfunding project.	Text
Country	The host country of the crowdfunding project	Text
Target	Amount of money the fundraiser targets.	Currency
Number of backers	Number of projects backed by creator.	Number
Average pledged	Average amount of money pledged over time.	Number
Creator's name	The name of the crowdfunding fundraiser	Text

Flexible fund	The option to receive funding amount even if funding target is not reached.	
Website	Availability of project website	Text
Twitter account	Fundraiser or project Twitter account.	Text
Facebook account	Fundraiser or project Facebook account.	Text
Youtube account	Fundraiser or project Youtube account.	Text

5. Method

Figure 2 gives a depiction of the generic set of steps that would be followed.

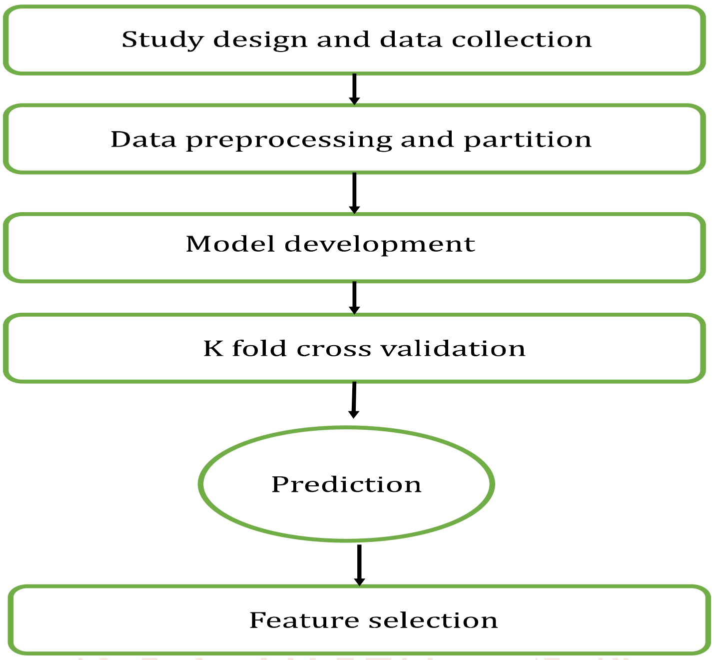

Figure 2Structural design of the research

5.1. Study design and data collection

The first step in the research after problem analysis is to determine what kind of data is appropriate for the research objectives and how much data is needed. The researcher must decide whether to use primary data or historical data and also choose the appropriate data collection strategy. In the context of cost, time, and paucity of data, this study collects historical data through existing statistics from various secondary sources. Advantages of using historical data collection for statistical analysis include low measurement error, availability of variables longitudinally or cross-sectionally, and the potential for replication (Vartanian, 2011). With respect to the design of the study, the current study adopts factorial designs whereby we focus on prediction in terms of finding the factors that affect the response variable.

5.2. Data pre-processing and exploratory data analysis

Data pre-processing refers to the addition, deletion, and transformation of data for the analysis. Data pre-processing is necessary to prepare the data to the type usable by the predictive or explanatory model. Also, pre-procession is necessary to take care of skewness and outliers in order to ensure high performance of the model. In particular, the most straightforward data transformation is centering and scaling which ensures that all the variables have zero means. In this study we shall consider the following data preprocessing objectives:

1. Omission of missing values: In the case of this study, we shall simply omit missing data values.
2. Data splitting: Data splitting is conducted to ensure that the models are useful for future instances. The first part of the split which makes about 70% to 80% proportion of the original data set and it is used to train the model for internal validation. The remaining proportion of the data serves as unseen data by the predictive model to use for the external validation. The simplest way to split data in machine learning is by taking a stratified random sampling from the data.

Exploratory data analysis (EDA) is a vital preliminary step in statistical modeling. EDA procedures could include summarizing the data numerically and graphically, reducing dimensionality, and a preparation for formal modeling. In statistical modeling EDA is applied without any strict rules that must be followed. Thus, each form of EDA method should only aim at assisting the purpose of capturing relationships that are possibly unknown or less formally expressed.

5.3. Model development

With predictive modeling the main priority is generating accurate predictions of new observations. Typically, there are single model-based algorithms to perform both explanatory and predictive tasks; however, ensemble algorithms are mainly utilized for predictive modeling. Most of these algorithms have their underlying statistical software packages with tunable parameters that influence the way the algorithms perform depending upon the data sets. The choice of an algorithm and a package also depends on the researcher's experience and the classification or regression approach.

5.4. Cross validation

Cross validation engages resampling techniques to conduct a multiple repetition of data resampling to produce an aggregated result using the training data. Cross validation is often used to tune model parameters to address the problem of overfitting for both predictive and explanatory modeling. Furthermore, cross-validation evaluation is important and efficient than just

observing the residuals of the model. This is because residual examination does not tell us how the model would perform on a data it has not already seen whereas the main purpose of cross validation is to provide an estimate for the performance of the model on a new data. There are several methods for choosing the portions of the sample to be used as a training and validation sets for cross-validation. The cross-validation technique used in this study is the repeated *k*-fold cross-validation. The reason for choosing the repeated *k*-fold cross-validation method is that the approach is less prone to selection bias and allows the training data set to be further randomly divided into *k* disjoint sets of equal size n/k, where *n* is the features in the data set. For each $k= 1,2,\cdots,K$ the cross-validation method fit the model with parameter λ to the other $K-1$ parts and computes it error in the prediction of the *k*th part. The process is repeated for many values of λ and selects value that makes the cross-validation (λ)smallest.

5.5. Model evaluation

We adopt a confusion matrix approach to evaluate the models. A Confusion matrix is the performance matrix used in the classification to determine the various performance related error types. We used the accuracy parameter of the confusion matrix to determine the classification efficiency of the models in this study.

5.6. Prediction

Ideally, prediction in supervised machine learning is conducted using the testing data to ensure that the model's performance is evaluated with samples that were not used to train the model in order to ensure that an unbiased sense of the model performance is ascertained.

5.7. Feature selection

Feature selection is integral to the predictive model to select the minimum required features to produce a valid model based on the training set. Feature importance gives the score of each feature toward the output variable in the model.

6. Results

Table 3 presents the data preprocessing and the frequency of both the response variables and the predictors. We can observe that 210 projects recorded a funding level that is between 0% to 20%, 12 projects recorded a funding level between 21% to 40%, 2 projects recorded a funding level between 41% to 60%, 3 projects recorded a funding level between 61% to 80%, 6 projects recorded a funding level between 81% to 100%, and 4 projects recorded a funding level above 101%. Also, with respect to the type of crowdfunding, 3 projects are donation-based crowdfunding, and 234 projects are rewards based crowdfunding projects. Similarly, 3 projects did not have the fundraiser's information and 234 projects produced the fundraiser's information with respect to the availability of websites, 226 projects did not have websites whereas 11 projects had websites. Again 191 projects where not publicized on social media whereas 46 projects were on social media.

Table 4 present the summary statistics of the numerical variables in the study. We can notice that the lowest duration for a project is one day and the maximum is 212 days. Also, the minimum number of backers for a project is zero whereas the maximum is 276. The mean average amount pledged is $27.16 and the maximum amount is $512.50.

Table 3

Variable		Preprocessing Value	Frequency
Response			
Percentage of funding	*FND*		
0 – 20		0	210
21 – 40		1	12
41 – 60		2	2
61 – 80		3	3
81 – 100		4	6
101+		5	4
Predictors			
Type of crowdfunding	*TYP*		
Donation		0	3
Rewards		1	234
Fundraiser's information	*INFO*		
No		0	3
Yes		1	234
Website	*WBS*		
No		0	226
Yes		1	11
Social Media	*SCM*		
No		0	191
Yes		1	46

Table 4

		Minimum	Median	Mean	Maximum
Predictors					
Duration	DRTN	1	55	44	212
Backers	BKS	0	0	6	276

Average pledge	AVP	0	0	27.16	512.50

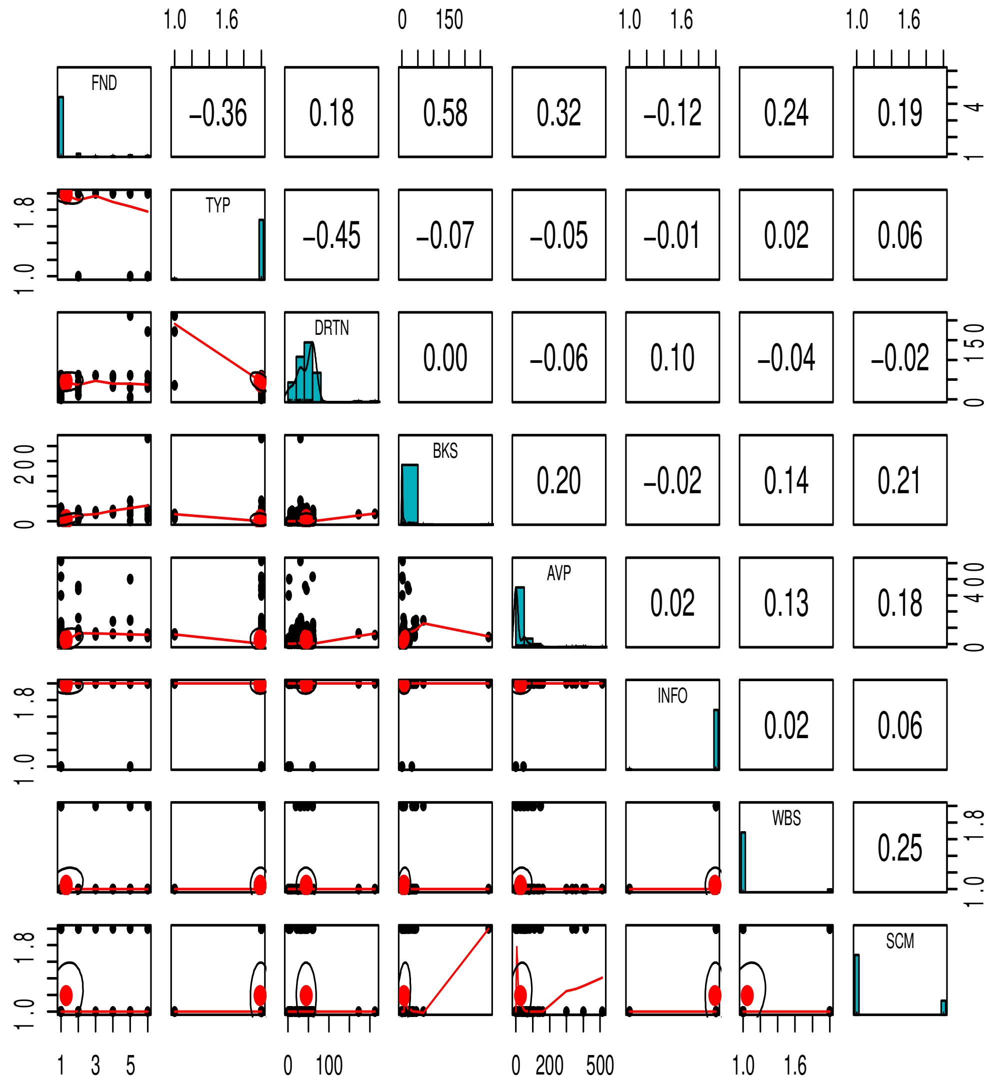

Figure 3 Correlation plot

Figure 3 presents the scatter plot of matrices for the variables used in the study. The lower off-diagonal shows the scatter plot, a regression line for a given pair of variables, and a diagonal histogram of each variable whereas the upper of diagonal shows the pairwise Pearson correlation. We can observe that the level of funding has strong correlation with the number of backers, giving us a first indication that the number of backers have a strong relationship with the level of funding.

6.1. Decision tree

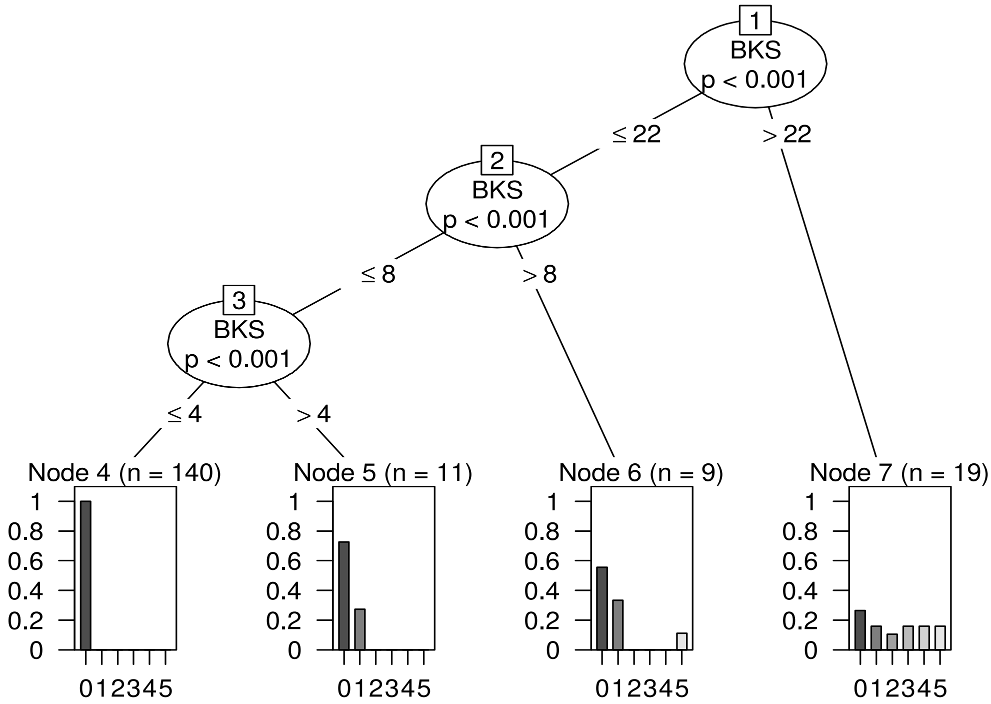

Figure 4 Decision tree diagram

Figure 4 shows the diagram for the decision tree classification model. It can be noticed that the number of backers for a given project (*BKS*) forms the root node of the representing the most important variable in the model. We can observe that there is 99% confidence level that funding for projects that have *BKS*>22 would be classified in node 7 where there is approximately 30% chance of realizing funding between 0 to 20%, Projects that have BKS≦22 would be decided by the number of backers that are 8 and below than 8 or above 8. Projects that have the number of backers > 8 are classified into node 6 where there is approximately 60% chance of achieving funding level between 0 to 20% and approximately 40% probability of achieving funding level between 21% to 40%. Projects that have the number of backers≦8 are decided by the number of backers ≦4 or > 4. Projects that have the number of backers > 4 are classified into node 5 where there is approximately 70% chance of attaining funding level between 0 to 20% and 30% chance of achieving funding level between 21% to 40%. Projects that have the number of backers ≦ 4 are classified into node 4 where there is a 100% chance of attaining funding between 0 to 20% funding level.

Table 5Actual and Predicted Confusion Matrix (Training data set)

		Predicted response category					
		0-20	21-40	41-60	61-80	81-100	100+
Actual response category	0-20	158	9	2	3	3	4
	21-40	0	0	0	0	0	0
	41-60	0	0	0	0	0	0
	61-80	0	0	0	0	0	0
	81-100	0	0	0	0	0	0
	100+	0	0	0	0	0	0

Table 5 shows the classification matrix of the predictions based on the decision tree algorithm using the train data. It can be noticed that the model made158 correct prediction and classification for actual projects that recorded 0% to 20% funding. However, there was no correct classification for the rest of the categories. The overall classification accuracy is approximately 88.26% indicating a prediction of the model.

Table 6 Actual and Predicted Confusion Matrix (Testing data set)

		Predicted response category					
		0-20	21-40	41-60	61-80	81-100	100+
Actual response category	0-20	52	3	0	0	3	0
	21-40	0	0	0	0	0	0
	41-60	0	0	0	0	0	0
	61-80	0	0	0	0	0	0
	81-100	0	0	0	0	0	0
	100+	0	0	0	0	0	0

Table 6 shows the classification matrix of the predictions based on the decision tree algorithm using the test data. It can be noticed that the model made 52 correct prediction and classification for actual projects that recorded 0% to 20% funding. The overall classification accuracy is approximately 89.65% indicating a prediction of the model.

6.2. Random Forest

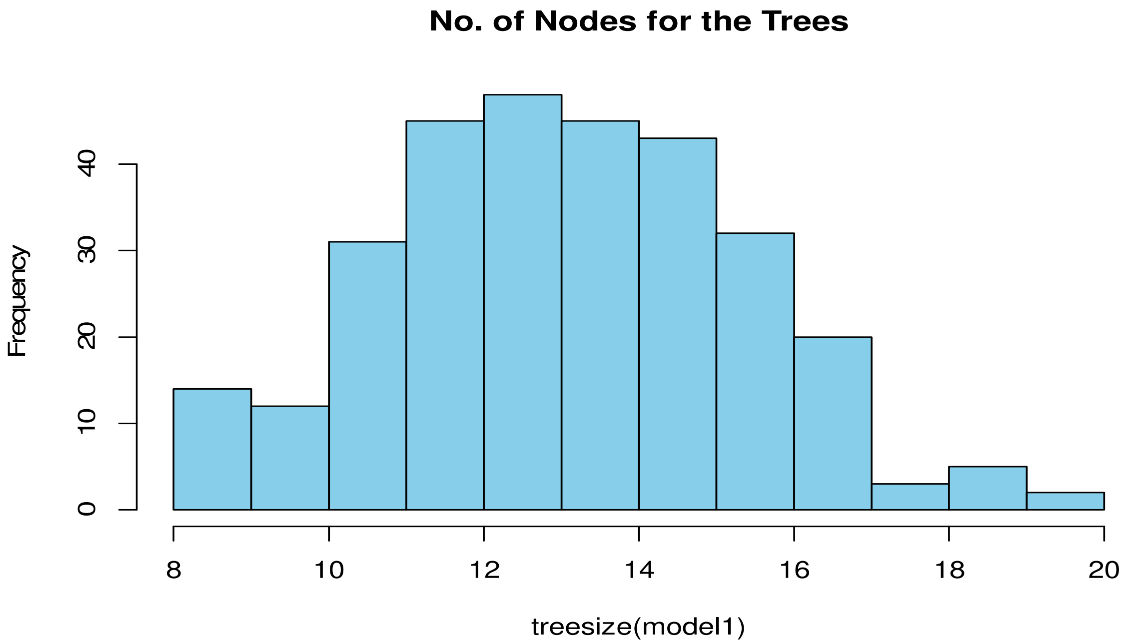

Figure 5Number of tree nodes

Figure 5 shows the optimum number of trees used in the random forest algorithm that produces the minimum out of bag (OOB) error rate to ensure that the model does not overfit. The model uses 12 trees with a frequency of 50.

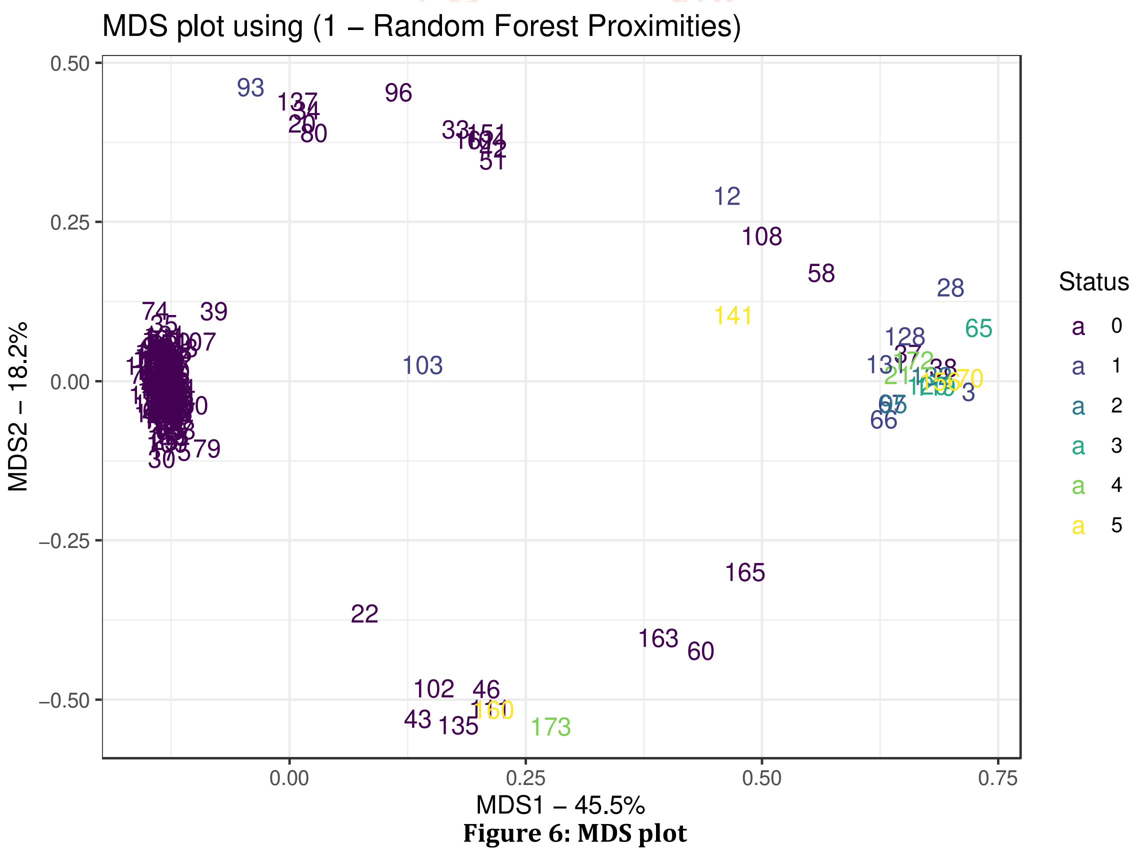

Figure 6: MDS plot

Figure 5 shows the multidimensional scaling (MDS) plot based on the proximity matrix of the random forest algorithm. The MDS plot is similar to the principal component analysis (PCA) bi plot to show the proximities of observations based on two-dimensional plot. Each point represents an observation and the between points can be used to identify subgroups of cases that keep together in the trees and also outliers that remain alone in the terminal node. It can be observed that the x-axis (MDS 1) accounts for 45.7% of the variations among the observations whiles the y-axis (MDS 2) accounts for 18.2% of the variations in the observations.

Table 7 Actual and Predicted Confusion Matrix (Training data set)

		Predicted response category					
		0-20	21-40	41-60	61-80	81-100	100+
Actual response category	0-20	158	0	0	0	0	0
	21-40	0	9	0	0	0	0
	41-60	0	0	2	0	0	0

	61-80	0	0	0	3	0	0
	81-100	0	0	0	0	3	1
	100+	0	0	0	0	0	4

Table 7 shows the classification matrix of the predictions based on the random forest algorithm using the training data. It can be observed that the model predicted and made a correct classification of 158 projects that obtained 0% to 20% of the campaign target. The model correctly predicted 9 projects that realized between 21% to 40% of the campaign target. The model correctly predicted 2 project that received between 41% to 60% of the campaign target. Again, correctly predicted 3 project that realized between 61% to 80% of the campaign target. Also, correctly predicted 3 projects that realized between 81% to 100% of the campaign target. Furthermore, correctly predicted 4 projects that realized over 100% of the campaign target. The overall classification accuracy based on the model using the test data is 100% indicating complete accuracy.

Table 8 Actual and Predicted Confusion Matrix (Testing data set)

		Predicted response category					
		0-20	21-40	41-60	61-80	81-100	100+
Actual response category	0-20	52	2	0	0	2	0
	21-40	0	1	0	0	0	0
	41-60	0	0	0	0	0	0
	61-80	0	0	0	0	0	0
	81-100	0	0	0	0	0	0
	100+	0	0	0	0	1	0

Table 8 displays the classification matrix of the predictions based on the random forest algorithm using the test data. We notice that the model predicted and made a precise classification of 52 projects that obtained 0% to 20% of the campaign target. The model correctly predicted 1 project that realized between 21% to 40% of the campaign target. The overall classification accuracy based on the model using the test data is 91. 37% indicating high classification accuracy.

Table 9 Variable importance

Variable	Mean decrease Gini
BKS	17.4
AVP	14.2
DRTN	5.42
WBS	1.16
SCM	0.487
INFO	0.359
TYP	0.164

Table 9 presents the variable importance according to mean decreasing Gini. We can notice that the number of backers is ranked as the most important variable followed by the average amount pledged.

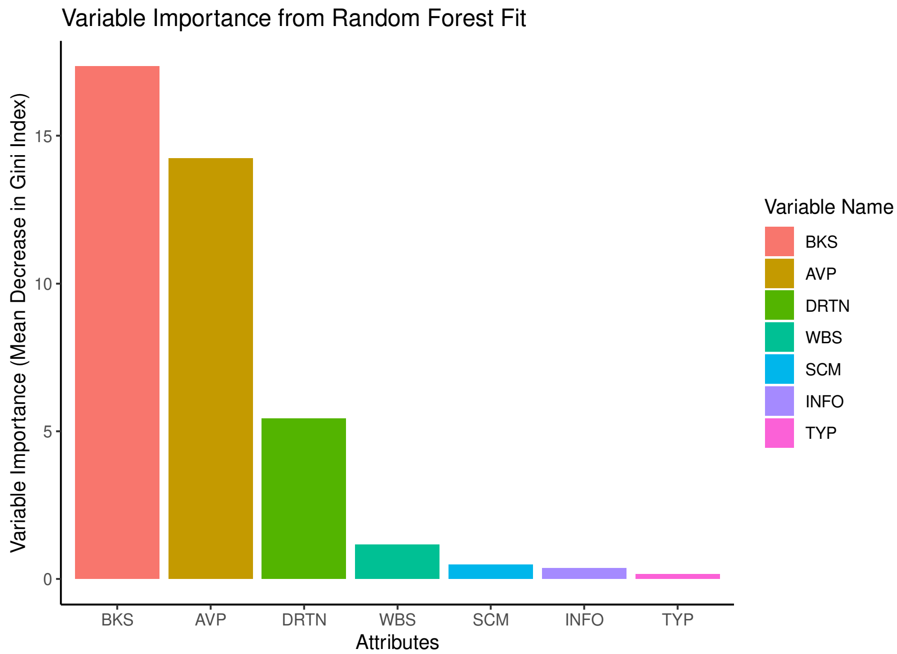

Figure 7 Random forest variable importance

Figure 7 presents the feature importance of the model based on the mean decrease Gini. It can be observed that the number of backers (*BKS*) have the highest value of approximately 17.4. We can also notice that the average pledged to the project (*AVP*) and the duration of the campaign (*DRTN*) have high values of approximately 14.2 and 5.42 respectively. Further, the use of website (*WBS*) has an index of 1.16, social media (*SCM*) has an index value of 0.487, provision of fundraiser's information (*INFO*) has an index of 0.359 and the type of crowdfunding (*TYP*) produced a value of 0.164.

6.3. Extreme Gradient Boosting

Figure 8 shows the simulation processes of the repeated cross-validation logloss error for the extreme gradient boosting algorithm where 7000 iterations are performed using both the train and test data. The 3058^{th} iteration produced the minimum logloss value of 0.150654 for train data and a minimum logloss value of 0.463064 for the test data.

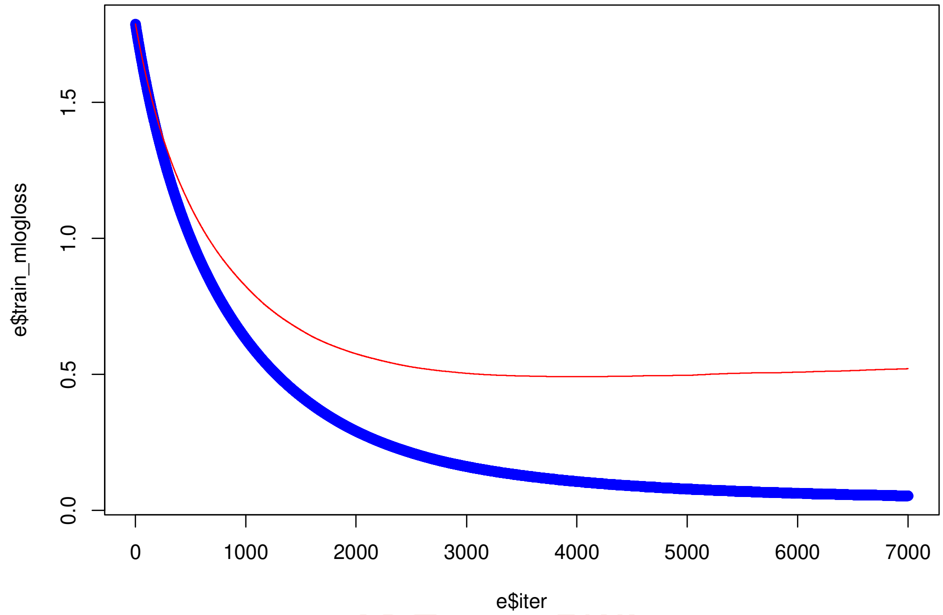

Figure 8 Extreme gradient boosting simulation

Table 10 Actual and Predicted Confusion Matrix (Training data set)

		Predicted response category					
		0-20	21-40	41-60	61-80	81-100	100+
Actual response category	0-20	158	0	0	0	0	0
	21-40	0	9	0	0	0	0
	41-60	0	0	2	0	0	0
	61-80	0	0	0	2	0	0
	81-100	0	0	0	0	3	0
	100+	0	0	0	0	0	4

Table 10 shows the classification accuracy for the predictions made by the extreme gradient boosting algorithm. It can be noticed that the model produced 158 correct predictions for actual funding between 0% to 20%, 9 correct predictions for funding between 21% to 40%, 2 correct prediction for funding amount recorded between 41% to 60%, 2 correct classification for funding between 61% to 80% of the campaign target, 3 correct classification for predictions of funding raise between 81% to 100% of the campaign target and 4 correct classification for predictions of over 100% funds raised from the campaign. The overall classification accuracy of the extreme gradient boosting algorithm for the train data is 100%.

Table 11 Actual and Predicted Confusion Matrix (Testing data set)

	Predicted response category			
		0-20	21-40	61-80
Actual response category	0-20	48	1	1
	21-40	2	2	1
	100+	2	0	1

Table 11 shows the actual and the predicted values classification matrix for the test data. It can be noticed that there is 48 correct classification for predictions of funding between 0% to 20% of the campaign target, 2 correct classification for predictions of funding between 21% to 40% of the campaign target, and 1 correct classification for predictions of funding between 61% to 80% of the campaign target. The overall extreme gradient boosting algorithm classification accuracy for the test data is 87.93%.

Table 12 Variable importance

Feature	Gain	Cover	Frequency
BKS	0.459227361	0.539407385	0.35420003
AVP	0.378494255	0.340286404	0.40584817
DRTN	0.151410339	0.102924857	0.20313001
SCM	0.007124189	0.009398867	0.02288778
WBS 1	0.003743856	0.007982487	0.01393401

Table 12 and Figure 9 presents the feature importance of the model on a scale of 0 to 1. It can be observed that the number of backers has the highest feature value of approximately 0.334 in the model indicating that Structural Social Capital has the highest incentive to contributors. It can also be observed that the average pledged to the project (*AVP*) and the duration of the

campaign (*DRTN*) feature as important variables in the model with values 0.32 and 0.20 respectively showing that Cognitive Social Capital influences funding decisions. Further, the platform used for the campaign (PLT) also offers the Relational Social Capital incentive to contributors with an importance value of approximately 0.73. Also, social media (*SCM*) representing Structural Social Capital has importance value of approximately 0.64 in the model.

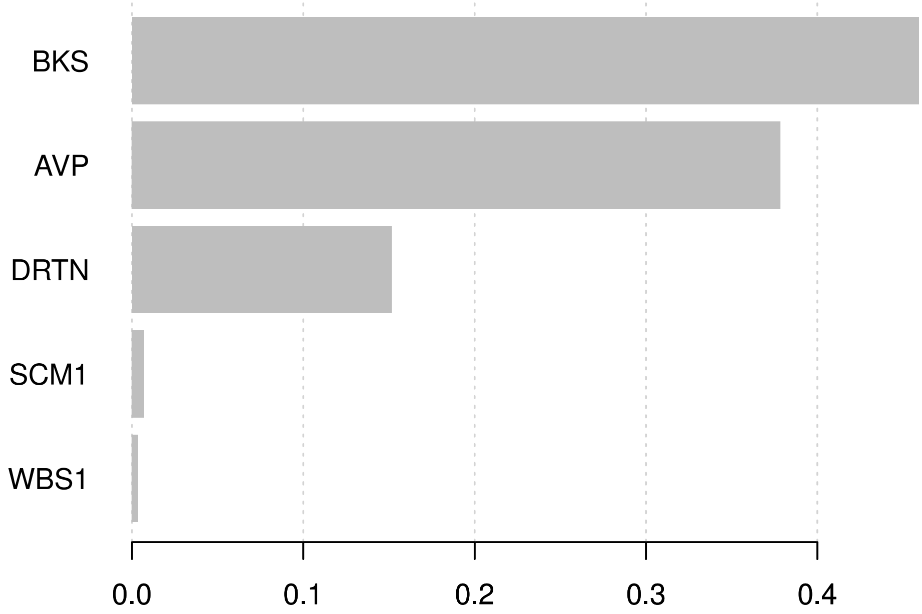

Figure 9 Variable importance

7. Conclusion

In this study we undertake a prediction modeling of the level of crowdfunding outcome with focus on African crowdfunding campaign. We observe that project attributes play a useful role in predicting the outcome of crowdfunding campaigns (Etter et al., 2014; Kamath and Kamat, 2016). Classification modeling of the outcome of crowdfunding campaigns has become an emerging research interest in recent years (Quercia and Crowcroft, 2014). Based on social capital theory, we conducted a supervised learning modeling of crowdfunding campaign outcome using different classifiers. Conducting classification of the level of funding using the training data, the ordinal logistic regression model produced a classification accuracy of 91.32%, the decision tree algorithm produced a classification accuracy of 88.16%, the random forest algorithm produced 100% classification accuracy, and the extreme gradient boosting algorithm produced100% classification accuracy. Thus, the ensemble models used in the study generated perfect prediction accuracies with the training data set. Using the testing data, the ordinal logistic regression yielded a classification accuracy of 87.5%, the decision tree algorithm produced a classification accuracy of 89.70%, the random forest algorithm generated a classification accuracy of 94.12%, whereas the extreme gradient boosting recorded a classification accuracy of 83.33%. Thus, the random forest model earned the highest classification accuracy performance.

Analyzing the important variables from the classifiers, irrespective of the model used, the number of backers is ranked as the most important variable for determining the level of funding for any given crowdfunding project. Also, the average pledged to the project is ranked second to the number of backers. The random forest algorithm and the extreme gradient boosting chose the duration of a project as the third most important variable in predicting the level of funding. Thus, the top three important predictors of the level of funding based on the study are the number of backers, the average pledged to the project, and the duration of a project.

References

[1] Agrawal, A., Catalini, C., & Goldfarb, A. (2015). Crowdfunding: Geography, Social Networks, and the Timing of Investment Decisions. *Journal of Economics & Management Strategy*, *24*(2), 253–274. https://doi.org/10.1111/jems.12093

[2] Ahmad, F. S., Tyagi, D., & Kaur, S. (2017). Predicting crowdfunding success with optimally weighted random forests. *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, 770–775. https://doi.org/10.1109/ICTUS.2017.8286110

[3] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[4] Burtch, G., Ghose, A., & Wattal, S. (2013). An Empirical Examination of the Antecedents and Consequences of Contribution Patterns in Crowd-Funded Markets. *Information Systems Research*, *24*(3), 499–519. JSTOR. https://www.jstor.org/stable/42004279

[5] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & implementation), Xgb. contributors (base Xgb. (2020). *xgboost: Extreme Gradient Boosting* (1.0.0.2) [Computer software]. https://CRAN.R-project.org/package=xgboost

[6] Cholakova, M., & Clarysse, B. (2015). Does the Possibility to Make Equity Investments in Crowdfunding Projects Crowd Out Reward–Based Investments?: *Entrepreneurship Theory and Practice*. https://doi.org/10.1111/etap.12139

[7] Choy, K., & Schlagwein, D. (2016). Crowdsourcing for a better world: On the relation between IT affordances and donor motivations in charitable crowdfunding. *Information Technology & People*,

29(1), 221–247. https://doi.org/10.1108/ITP-09-2014-0215

[8] Chung, J., & Lee, K. (2015). A Long-Term Study of a Crowdfunding Platform: Predicting Project Success and Fundraising Amount. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 211–220. https://doi.org/10.1145/2700171.2791045

[9] Ciuchta, M., Letwin, C., Stevenson, R., & McMahon, S. (2016). Regulatory Focus and Information Cues in a Crowdfunding Context. *Applied Psychology*, *65*, n/a-n/a. https://doi.org/10.1111/apps.12063

[10] Cumming, D., Leboeuf, G., & Schwienbacher, A. (2014). Crowdfunding Models: Keep-it-All vs. All-or-Nothing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2447567

[11] Davies, W. E., & Giovannetti, E. (2018). Signalling experience & reciprocity to temper asymmetric information in crowdfunding evidence from 10,000 projects. *Technological Forecasting and Social Change*, *133*(C), 118–131. https://econpapers.repec.org/article/eeetefoso/v_3a133_3ay_3a2018_3ai_3ac_3ap_3a118-131.htm

[12] Farnel, M. (2014). Kickstarting trans*: The crowdfunding of gender/sexual reassignment surgeries: *New Media & Society*. https://doi.org/10.1177/1461444814558911

[13] Galuszka, P., & Brzozowska, B. (2017). Crowdfunding and the democratization of the music market. *Media, Culture & Society*, *39*(6), 833–849. https://doi.org/10.1177/0163443716674364

[14] Gerber, E., & Hui, J. (2014). Crowdfunding: Motivations and Deterrents for Participation. *ACM Transactions on Computer-Human Interaction*, *20*, 34–32.

[15] KPMG. (2015). *Sector report: Banking in Africa 2015. Johannesburg, South Africa: KPMG.* http://www.kpmg.com/Africa/en/IssuesAndInsights/Articles-Publications/General-IndustriesPublications/Pages/Banking-in-Africa-2015.aspx.

[16] Lam, P. T. I., & Law, A. O. K. (2016). Crowdfunding for renewable and sustainable energy projects: An exploratory case study approach. *Renewable and Sustainable Energy Reviews*, *60*(C), 11–20. https://econpapers.repec.org/article/eeerensus/v_3a60_3ay_3a2016_3ai_3ac_3ap_3a11-20.htm

[17] Lin, M., & Viswanathan, S. (2016). Home Bias in Online Investments: An Empirical Study of an Online Crowdfunding Market. *Management Science*, *62*(5), 1393–1414. https://econpapers.repec.org/article/inmormnsc/v_3a62_3ay_3a2016_3ai_3a5_3ap_3a1393-1414.htm

[18] Lopez de Prado, M. (2018). *Advances in Financial Machine Learning (Chapter 1)* (SSRN Scholarly Paper ID 3104847). Social Science Research Network. https://papers.ssrn.com/abstract=3104847

[19] Meer, J. (2014). Effects of the price of charitable giving: Evidence from an online crowdfunding platform. *Journal of Economic Behavior & Organization*, *103*(C), 113–124. https://ideas.repec.org/a/eee/jeborg/v103y2014icp113-124.html

[20] Mollick, E. (2014). The Dynamics of Crowdfunding: An Exploratory Study. *Journal of Business Venturing*, *29*, 1–16. https://doi.org/10.1016/j.jbusvent.2013.06.005

[21] Ordanini, A., Miceli, L., Pizzetti, M., & Parasuraman, A. (2011). Crowd-funding: Transforming customers into investors through innovative service platforms. *Journal of Service Management*, *22*(4).

[22] Parker, S. C. (2014). Crowdfunding, cascades and informed investors. *Economics Letters*, *125*(3), 432–435. https://doi.org/10.1016/j.econlet.2014.10.001

[23] Ren, X., Xu, L., Zhao, T., Zhu, C., Guo, J., & Chen, E. (2018). Tracking and Forecasting Dynamics in Crowdfunding: A Basis-Synthesis Approach. *2018 IEEE International Conference on Data Mining (ICDM)*, 1212–1217. https://doi.org/10.1109/ICDM.2018.00161

[24] Ryu, S., & Kim, Y.-G. (2015). *A Typology of Crowdfunding Sponsors: Birds of a Feather Flock Together?* (SSRN Scholarly Paper ID 2734097). Social Science Research Network. https://papers.ssrn.com/abstract=2734097

[25] SONG, Y., & LU, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

[26] The World Bank Group. (2015). *Crowdfunding in Emerging Markets: Lessons from East African Startups.* The World Bank Group.

[27] Vartanian, T. (2011). Secondary Data Analysis. *Secondary Data Analysis*, 1–216. https://doi.org/10.1093/acprof:oso/9780195388817.001.0001

[28] Yu, P.-F., Huang, F., Yang, C., Liu, Y.-H., Li, Z., & Tsai, C.-R. (2018). Prediction of Crowdfunding Project Success with Deep Learning. *2018 IEEE 15th International Conference on E-Business Engineering (ICEBE)*. https://doi.org/10.1109/ICEBE.2018.00012

[29] Yuan, H., Lau, R., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, *91*. https://doi.org/10.1016/j.dss.2016.08.001

[30] Zheng, H., Li, D., Wu, J., & Xu, Y. (2014). The role of multidimensional social capital in crowdfunding: A comparative study in China and US. *Information & Management*, *51*(4), 488–496. https://doi.org/10.1016/j.im.2014.03.003